DOCUMENT RESUME

ED 353 267 TM 018 969

AUTHOR Tesh, Anita S.; And Others
TITLE Evaluation of the Effectiveness of Alternate Questionnaire Item Response Options.
PUB DATE Feb 92
NOTE 16p.; Paper presented at the Annual Meeting of the North Carolina Association for Research in Education (Chapel Hill, NC, February 20, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Effective Schools Research; *Elementary School Teachers; Elementary Secondary Education; Item Response Theory; *Likert Scales; Public School Teachers; *Questionnaires; *School Effectiveness; *Secondary School Teachers; Test Construction; Test Format; *Test Reliability

ABSTRACT

The effectiveness of two five-point Likert response scales with different verbal anchors was compared. Four anonymous written questionnaires addressing characteristics of effective schools were developed. Each questionnaire was prepared in two formats, one with each response scale. The 4 questionnaires were composed of between 140 and 200 items. A balance of positive and negative items was used. Each of 539 public school teachers completed both formats, with the order of administration randomized. The two response scales were found to be similar in terms of average required completion times, reliabilities, and respondents' written comments, but they differ in non-response rates. These results suggest a potential series of investigations that instrument developers may wish to use in developing questionnaires. It appears feasible and worthwhile to include comparison of alternate response scales in the pilot survey. Seven graphs present study data, and there is a 17-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

ANITA S. TESH

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED353267

Evaluation of the Effectiveness of Alternate Questionnaire Item Response Options

Anita S. Tesh
School of Nursing
University of North Carolina at Greensboro
Greensboro, NC 27412-5001

Claudia S. McKenzie
Department of Mathematics
Central Piedmont Community College
Charlotte, NC

Richard M. Jaeger
Center for Educational Research and Evaluation
University of North Carolina at Greensboro
Greensboro, NC 27412-5001

Paper presented at the Annual Meeting of the
North Carolina Association for Research in Educational
Chapel Hill, NC., 20 February 1992

TM018916

BEST COPY AVAILABLE

Evaluation of the Effectiveness of Alternate Questionnaire Item Response Options

Abstract

In this study the effectiveness of two five-point Likert response scales with different verbal anchors was compared. Four anonymous written questionnaires addressing characteristics of effective schools were developed. Each questionnaire was prepared in two formats: one with each response scale. Each of 539 respondents completed two questionnaires: one in each response-scale format. Order of administration was randomized. The two response scales were found to be similar in terms of average required completion times, reliabilities, and respondents' written comments, but to differ in non-response rate.

Evaluation of the Effectiveness of Alternate Questionnaire Item Response Options

Objectives

In the course of development of a comprehensive school evaluation system for state-wide use, the authors drew upon both the literature on the characteristics of effective schools, and the literature on the assessment of valued student outcomes. As a part of the comprehensive evaluation system, written survey questionnaires based upon the characteristics of effective schools were developed for certified and noncertified school personnel. These questionnaires were intended to assess the degree to which respondent's felt their school exemplified the characteristics of effective schools.

The objective of this study was to compare the effectiveness of alternate response options (i.e., answer scales) in eliciting information from respondents. The effectiveness of two five-point Likert scales was compared. One scale had extremes labeled "strongly agree" and "strongly disagree;" the other had extremes labeled "true" and "false."

Perspective

Cronbach (1970) distinguished tests designed to measure respondent's abilities, aptitudes, or achievement (termed tests of *maximum performance)* from tests designed to measure interests, opinions, attitudes, motives, or personality traits (termed tests of *typical performance*). The items of maximum performance tests generally have a single correct answer, while the items of tests of typical performance do not have correct or incorrect answers. Summated rating scales, often called Likert scales after the original developer, have historically been among the most frequently used methods in tests of opinion and attitude (Aiken, 1973, 1985; Anastasi, 1988; Burns & Grove, 1987; Nunnaly, 1978; Rossi, Wright, & Anderson, 1983).

A summated rating scale is composed of a series of statements which reflect favorably or unfavorably on the attitude object (here, the degree to which the school exemplifies various characteristics of effective schools). A response scale is given for each statement. After each respondent has indicated his response to each statement, the scores are summed and the respondent's opinion represented by the total. In constructing the scales, it is generally recommended that both positive and negative statements be used; scoring is reversed for negative statements. (Rossi, et al., 1983).

The original Likert response scale was a five point scale with responses labeled "strongly agree," "agree," "undecided," "disagree," and "strongly disagree" (Aiken, 1973; Anastasi, 1988; Rossi, et al., 1983). This five-point response scale is still commonly used, however, some test-developers use scales with more, or less, than five points; and some use labels other than "strongly agree," "agree," "undecided," "disagree," and "strongly disagree." In some cases, all intermediate points on the scale are labeled; in others, they are not. Frisbie and Brandenburg (1979), and Lam and Klockars (1982) point out that, in questionnaire development, considerable time and effort is typically spent on the selection of item stems, while little attention is paid to selection of the response scale for the questionnaire items.

Literature exists on the selection of the number of points on a response scale, and on the advantages of odd- and even-numbered scales (c.f. Burns & Grove, 1987; Comfrey, 1988; Grady & Wallston, 1988; Rossi, et al., 1983). Scales with more response points are advocated as increasing the variance of scores, and allowing for finer distinctions (Comfrey, 1988; Kerlinger, 1973; Nunnaly, 1978). However, longer scales are also more cumbersome (Nunnaly, 1978; Edelbrock, 1983). Nunnaly (1978) suggests that, while reliability typically increases with each additional scale step up to 20, in some cases the longer scales will irritate or confuse respondents, causing them to omit answers to many questions.

Scales with even numbers of response points are thought by some to force respondents to state an opinion or make a decision, thus increasing score variance and avoiding a midpoint "response set." Scales with a midpoint are argued by others to allow for the expression of valid "neutral" positions (Nunnaly, 1978). There is no consensus on the best number of points for a response scale, and Grady & Wallston (1988) conclude that the choice between five-, six,- or seven-point scales is fairly arbitrary, and has little effect on scale reliability or validity. After examining various empirical and psychometric arguments, Nunnaly (1978, p. 596) concludes that when responses are summed over multiple items, the choice of even- or odd-numbered response scales "is not very important," and may be left largely to the judgment of the experimenter.

Lam and Klockars (1982), and Frisbie and Brandenberg (1979) compared response scales with all intermediate points labeled to response scales with only the endpoints labeled. Both sets of investigators concluded that in some circumstances labeled intermediate points are preferable to unlabeled intermediate points, however, scales with unlabeled intermediate points produced results very similar to scales with all points labeled. Frisbie and Brandenberg

(1979) found no difference between scales with points designated by letters or by numbers.

If the traditional "strongly agree/strongly disagree" response scale labels are not automatically adopted, there is little empirical evidence to guide the instrument developer in the selection of the labels for the points on the response scale. The traditional response scale has been criticized as being imprecise (Bernstein, Borkovec, & Coles, 1986). Variations of the traditional response scale are in wide use, including scales with labels representing such underlying constructs as frequency (i.e.,"none of the time" to "all of the time"), certainty (i.e., "definitely" to "definitely not"), and quality (i.e., "poor" to "excellent") (Allen & Yen, 1979; Bernstein, Borkovec, & Coles, 1986; Carson, 1986; Comfrey, 1988; Nunnaly, 1978). French-Lazovik and Gibson (1984), and Lam and Klonkars (1982) found that the degree of skew of the distribution of responses was in part a function of the verbal labels attached to scale midpoints. When intermediate points were given more positive (or favorable) labels, the distribution of scores was altered from those obtained with the same item stems and response scale endpoints, but with less positive labels for the intermediate points. Neither set of researchers compared the results of equally spaced response scales with different labels on all points. Comparisons of the effectiveness of alternate response scales to the traditional scale (or to each other) are not generally available.

Methods

Instruments and Administration

The first step in the construction of a summated rating scale is the selection of the item statements (i.e., the "item stems"). Description of the exact method by which the questionnaire items used in this study were selected is beyond the scope (and beside the point) of this paper. In brief: items were derived from review of the professional literature and other instruments designed to assess the characteristics of effective schools. Final item selection was based on multiple panel reviews, and assessment of the psychometric properties (e.g., factor structure, point-biserial correlations) of the items and instruments.

Four anonymous, written questionnaires addressing the characteristics of effective schools were developed for teachers. Each of these four questionnaires was composed of between 140 and 200 items. A balance of positively and negatively stated items was used. A typical item stem might be: "This school provides me sufficient freedom to set priorities for my own work." The items of each questionnaire composed seven to eight subscales.

Each questionnaire was prepared in two forms: one with a five-point "true/false" response scale, the other with a five-point "agree/disagree" response scale. On the "true/false" response scale, the points were labeled "false," "more false than true," "equally true and false," "more true than false," and "true." On the "agree/disagree" response scale, the points were labeled: "disagree," "disagree somewhat," "neither agree nor disagree," "agree somewhat," and "agree." Thus, midpoints were neutral and intermediate points evenly spaced on both scales.

At each pilot-study school, questionnaires and response formats were spiraled so that each teacher took two of the four questionnaires, one in the "agree/disagree" format, and the other in the "true/false" format. The order in which teachers took the questionnaires and the response formats was randomized. Respondents were directed to record the beginning and completion times for each questionnaire they took.

Analytic Procedures

In deciding to empirically compare the effectiveness of two response scales in eliciting information from respondents, we felt that several factors should be considered. The time required by respondents to complete an instrument with a given response scale should be considered, as well as respondents' reactions to the scales. The psychometric properties of an instrument with each response scale must be compared. The nonresponse rates to instruments with each response scale must be compared, and, finally, respondents' reactions to the response scales should be taken into consideration. One desires the final instrument (with the selected response scale) to have sound psychometric properties and a low item nonresponse rate, to take as little time as possible to complete, and to be acceptable to respondents.

A variety of analytic procedures were carried out to satisfy the objective stated above. Respondents' written comments regarding the two response scale formats were solicited and summarized. Univariate statistics on the time required to complete questionnaires in each format were calculated and compared. Coefficient alpha reliabilities for questionnaire subscales in each response format were also estimated and compared. (In calculating coefficient alphas, the pooled correlation matrices of within-school responses to items were used.) Finally, item nonresponse rates were accumulated within subscales and compared across response scale formats.

Data Source

Twelve public schools (four elementary, three middle, and five secondary schools) from a southern state participated in a pilot-study to compare the two questionnaire response scales. These schools were selected to represent the spectrum of: 1) geographic locations within the state, 2) urban versus rural character, and 3) level of affluence. At each pilot-study school, every employed teacher completed a questionnaire with the "agree/disagree" response scale and one with the "true/false" response scale. A total of 539 teachers completed questionnaires.

Results and Implications

Respondents' primary comment regarding the response scales was to indicate the desire for a "don't know" option. A total of 91 teachers indicated a desire for such an option. Respondents made relatively few comments regarding the labels for the two response scales under investigation. Two teachers stated that they preferred the "agree/disagree" response scale to the "true/false" response scale. Two other teachers indicated that they felt the labels on the "true/false" scale (e.g., "more true than false") were not appropriate. Another teacher suggested a continuum of "observed" to "not observed" rather than either of the response scales investigated. No clear pattern of preference for response scale can be inferred from these few comments.

Between 87 and 131 teachers provided data on the time required to complete each questionnaire in each response scale format. Univariate statistics on the time required by respondents to complete each questionnaire form in each response scale format were calculated. T-tests on average completion times did not detect significant differences (at $\alpha = 0.05$) between the two response scales for any of the four questionnaires. When the null hypothesis: H_O: $Time_{a/d} = Time_{t/f}$ was tested for each questionnaire, the absolute values of the t-statistics were between 0.44 and 1.11, with associated probability levels of 0.66 to 0.27. The difference between the average time required to complete a questionnaire in "agree/disagree" format and in "true/false" format was less than two minutes in all four cases. In two cases, the difference was less than one minute. Mean and median completion times for the two response scales are presented graphically in Tables 1 and 2 on the following page.

Table 1: Average Completion Times by Response Scale Format

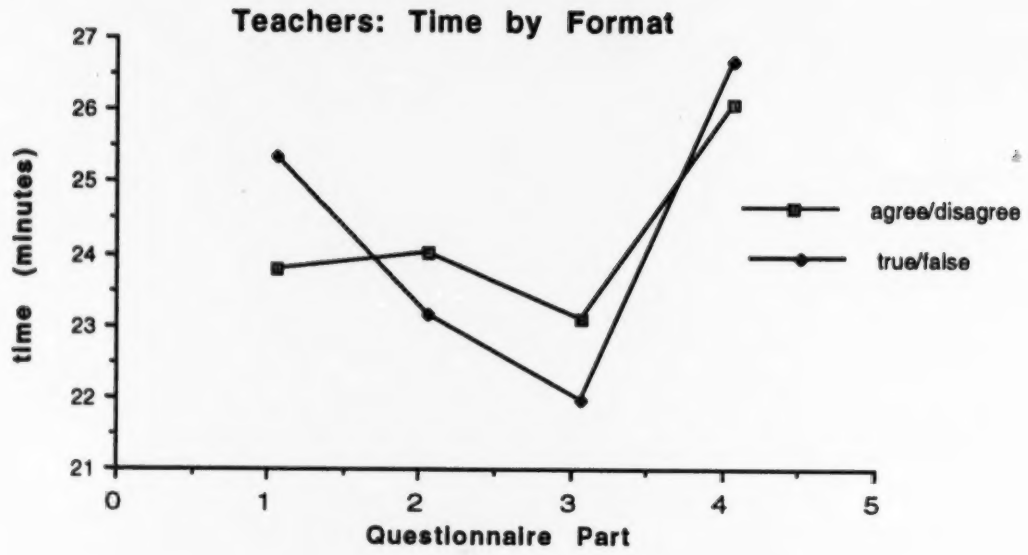
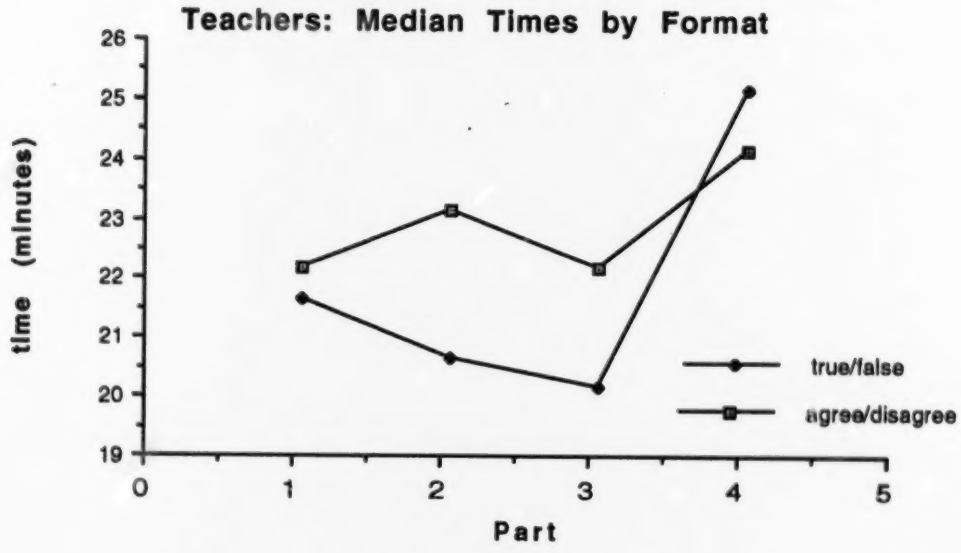


Table 2: Median Completion Times by Response Scale Format



For each of the four questionnaires, the standard deviation of the time required to complete the questionnaire was greater for the "true/false" format than for the "agree/disagree" format. When tested, the difference between the variances was statistically significant (at $\alpha = 0.05$) for only one of the four questionnaires.

To compare the estimated reliability of scales in the true/false and agree/disagree response formats, Cronbach's alpha (internal-consistency reliability) was calculated separately for the two response scale formats for each initial, unrevised subscale of the questionnaires. These values of Cronbach's alpha are presented graphically in Table 3, below. As can be seen from Table 3, comparison of preliminary values of Cronbach's alpha suggested no superiority of one response format over the other. A similar pattern was seen when the contribution of each item to subscale reliability was compared for items in the two response formats.

Table 3

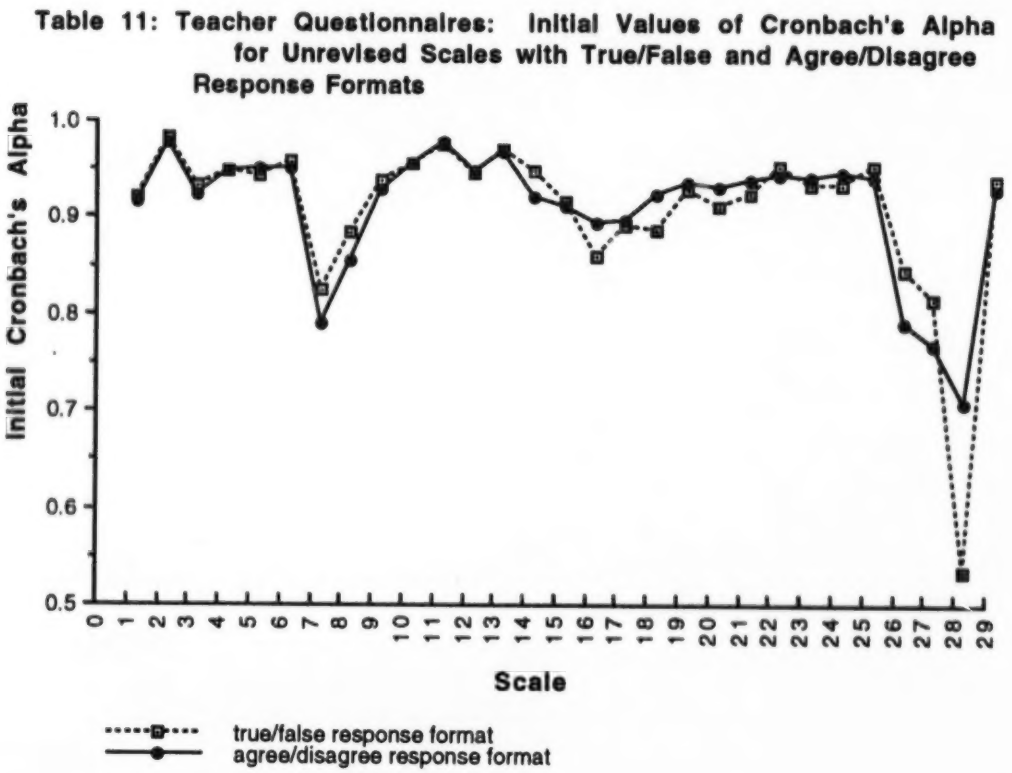


Table 11: Teacher Questionnaires: Initial Values of Cronbach's Alpha for Unrevised Scales with True/False and Agree/Disagree Response Formats
Initial Cronbach's Alpha
1.0
0.9
0.8
0.7
0.6
0.5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Scale
true/false response format
agree/disagree response format

The two response scales were also compared in terms of the nonresponse rates elicited by questionnaire subscales with each response scale. Between 109 and 139 respondents provided data necessary to calculate subscale nonresponse rates for each questionnaire. Nonresponse rates were calculated separately for the two response scale formats for the subscales of each questionnaire. These nonresponse rates are presented graphically in Tables 4 through 7 below. (Nonresponse rates for subscales of Questionnaire 1 are shown in Table 4; nonresponse rates for subscales of Questionnaires 2 through 4 are shown in Tables 5 through 7).

As can be seen from Tables 4 through 7, on average, the nonresponse rates for the questionnaires were substantially greater for the "true/false" response scale than for the "agree/disagree" response scale.

Table 4: Nonresponse Rate by Response Scale Format for Subscales of Questionnaire 1

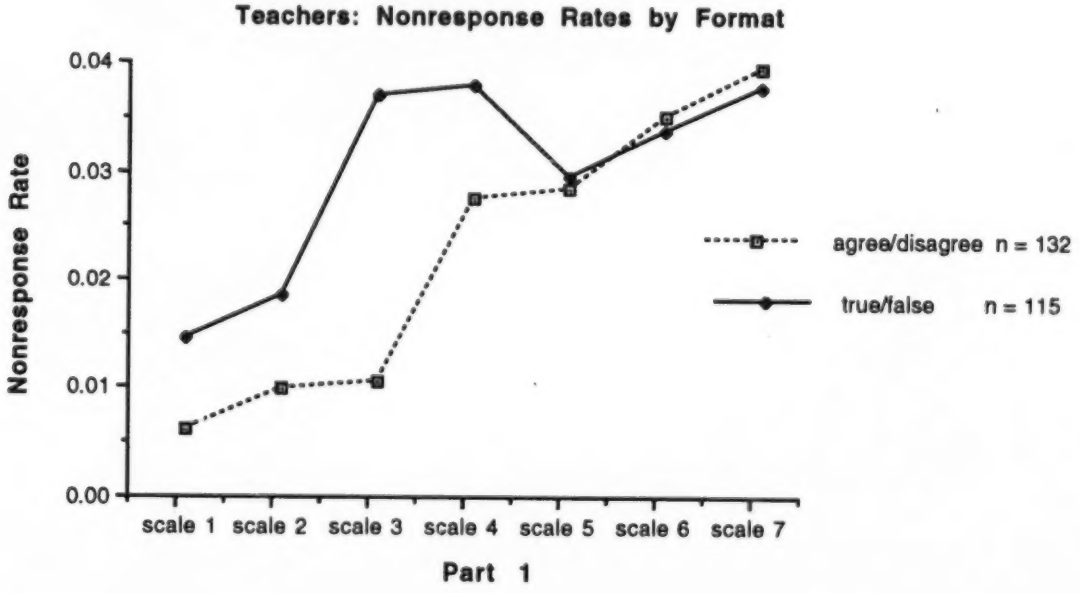


Table 5: Nonresponse Rate by Response Scale Format for Subscales of Questionnaire 2

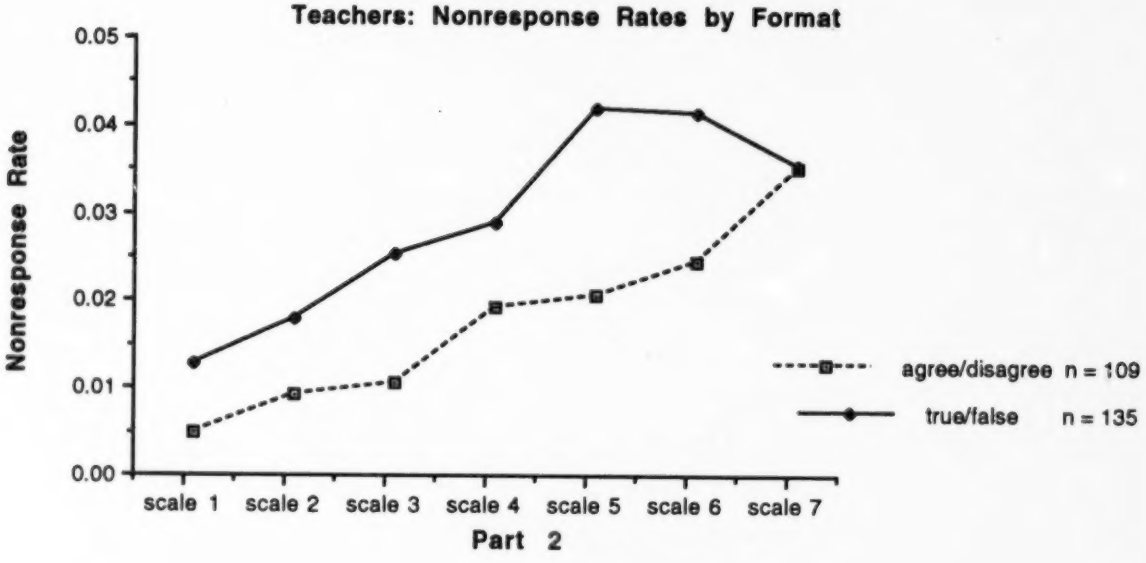


Table 6: Nonresponse Rate by Response Scale Format for Subscales of Questionnaire 3

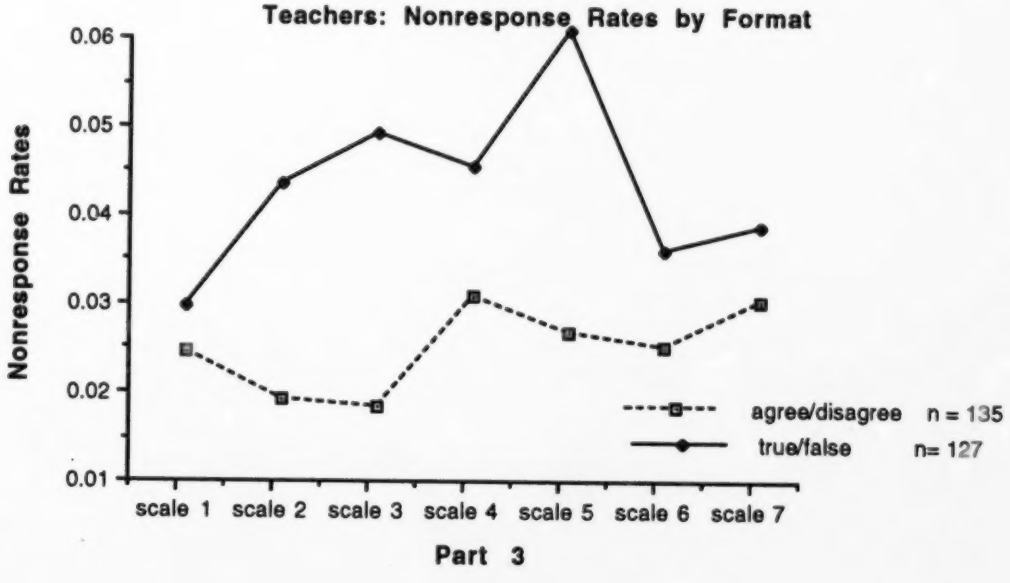
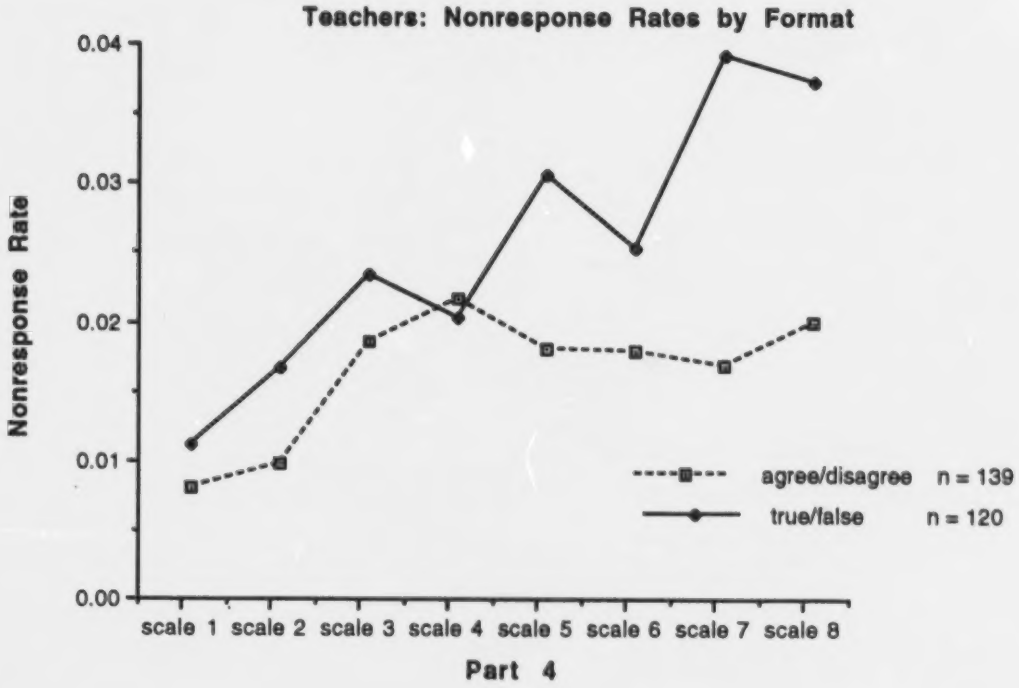


Table 7: Nonresponse Rate by Response Scale Format for Subscales of Questionnaire 4



Summary and Conclusion

In general, the two response scale formats investigated in this study were found to be similar in terms of respondents' written comments, average required completion times and questionnaire subscale reliabilities. However, the item nonresponse-rate was substantially greater for questionnaires with the "true/false" response scale than for the same questionnaires with the "agree/disagree" response scale. It was determined that, for these questionnaires, the "agree/disagree" response option was more effective in eliciting information from respondents, and this format was selected and developed for the field-study instruments.

This study also suggests a potential series of investigations which instrument developers may wish to consider when developing questionnaires. Instrument development typically includes a pilot-study phase to select item stems. The results of this study suggest that it is feasible and worthwhile to include comparison of alternative

response scales in the pilot study. If alternative response scales are to be compared, the authors suggest that investigators gather comparative data on: 1) completion times, 2) questionnaire or subscale reliabilities, 3) nonresponse rates, and 4) respondents' reactions, presented as written and/or verbal feedback. Investigators may also wish to gather comparative data on the distributional properties of respondents' total scores, investigation of which was precluded in this study by situational constraints.

References

Aiken, L.R. (1985). *Psychological Testing and Assessment.* Boston: Allyn & Bacon.

Aiken, L.R. (1973). *Psychological and Educational Testing.* Boston: Allyn & Bacon.

Allen, M.J., & Yen, W.M. (1979). *Introduction to Measurement Theory.* Montery, CA: Brooks/Cole Publishing Company.

Anastasi, A. (1988). *Psychological Testing.* 5th ed. New York: McMillan.

Bernstein, D.A.; Borkovec, T.D.; & Coles, M.G. (1986). Assessment of anxiety. In A.R. Ciminero, K.S. Calhoun, & H.E. Adams (eds.) *Handbook of Behavioral Assessment.* New York: John Wiley & Sons.

Burns, N., & Grove, S. (1987) *The Practice of Nursing Research: Conduct, Critique, and Utilization.* Philadelphia: W.B. Saunders.

Carson, T.P. (1986). Assessment of depression. In A.R. Ciminero, K.S. Calhoun, & H.E. Adams (eds.) *Handbook of Behavioral Assessment.* New York: John Wiley & Sons.

Comrey, A.L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56(5),* 754-761.

Cronbach, L.J. (1970). *Essentials of Psychological Testing.* 3rd ed. New York: Harper & Rowe.

Edelbrock, C. (1983). Problems and issues in using rating scales to assess child personality and psychopathology. *School Psychology Review, 12(3),* 293-299.

French-Lazovik, G., & Gibson, C.L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement, 8(1),* 49-57.

Frisbie, D.A., & Brandenburg, D.C. (1979). Equivalence of questionnaire items with varying response formats. *Journal of Educational Measurement, 16,* 43-48.

Grady, K.E., & Wallston, B.S. (1988). *Research in Health Care Settings.* London: Sage Publications.

Kerlinger, F.N. (1973). *Foundations of Behavioral Research.* 2nd ed. New York: Holt, Rinehart, and Winston.

Lam, T.C. & Klockars, A.J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement, 19(4),* 317-322.

Nunnaly, J.C. (1978). *Psychometric Theory.* 2nd ed. New York: McGraw-Hill.

Rossi, P. H.; Wright, J.D. & Anderson, A.B. (1983). *Handbook of Survey Research.* New York: Academic Press.